DR. JENNIFER KRETCHMAR is the senior assistant director of research in the Office of Undergraduate Admissions at the University of North Carolina at Chapel Hill. She earned her B.A. in psychology at Pennsyvania State University (PA) and her doctorate in education at University of North Carolina at Chapel Hill.

By Dr. Jennifer Kretchmar

Assessing the Reliability of Ratings

Used in Undergraduate Admission Decisions

Many colleges and universities receive thousands of applications for freshman admission every year. To facilitate the process of evaluating each and every applicant in a relatively short amount of time, schools often devise quantitative ratings scales to summarize student characteristics. The ratings give readers a shorthand way to communicate the qualities of each student, and sometimes play a critical role in determining whether or not a student is offered admission.

Given the significance of the decision for both the student and the institution to which he or she is applying, admission officers should have a vested interest in the reliability of their ratings scales and should be asking if students would receive the same ratings if their applications were reviewed by different readers. Ironically, even as admission officers evaluate applicants according to strict standards, they sometimes fail to evaluate their own practices with similar precision.

Often reliability among raters (interrater reliability) is measured using simple correlations or kappa coefficients (Cohen, 1960). Both methods, however, have limitations. Correlation measures the degree to which two raters co-vary instead of the level of agreement; a rater who gives consistently high ratings and a rater who gives consistently low ratings might produce a high correlation, even though none of their ratings are the same (Hux et al., 1997). The kappa coefficient can be used to determine whether agreement between raters exceeds chance levels of agreement, but it doesn't further qualify levels of agreement (Uebersax, 2003).

In short, for an accurate rating, admission professionals must determine different sources of variations in scores. Generalizability theory allows for this consideration and, therefore, has several advantages over more traditional measures of reliability. Partitioning variance according to its source is referred to as the G study. Depending on the situation, variation in scores might be due to raters (e.g., raters might have personal biases), questions (e.g., students might respond to one essay prompt better than another) or the number of testing occasions (e.g., students take a test multiple times). All of these sources of variation reflect error. Ideally, most of the variation in scores should reflect true differences among the individuals being rated. For example, a student who receives a higher essay score than another student should demonstrate better writing skills. By performing what are called "Decision studies" or "D studies," generalizability theory helps you determine how to maximize true differences and minimize error. Finally, generalizability theory also allows reliability estimates to be tailored to the anticipated use of the scores—for example, whether or not the scores will be used to make relative or absolute decisions about students (Swartz, et al., 1999).

The purpose of the following research was to use generalizability theory to investigate the reliability of six ratings scores used when evaluating applicants for admission to a university's undergraduate program. Reliability estimates were generated for these scales for both relative and absolute decisions—for the purpose of rating an individual applicant relative to all other applicants, and for the purpose of making a decision to offer or decline admission.

Method

Subjects

The subjects in this study were applicants to a prominent Southeastern public university. All subjects applied for undergraduate admission for Fall 2003. A random sample of 50 applicants was selected from the entire pool of over 17,000. In general, the academic characteristics of the sample mirrored those of the overall population; the average SAT of the sample was 1264; 54 percent were ranked in the top 10 percent of their high school class; and their average GPA was 4.01. The demographics of the sample varied slightly from those of the population; out-of-state applicants were underrepresented in the sample while female applicants were overrepresented. The sample mirrored the population in terms of children of alumni and traditionally underrepresented minorities—African-American, American-Indian and Latino.

Ratings Scales

Each application was rated by each reader in six different areas—academic program, academic performance, extracurricular activities, essays, potential to contribute, and potential to benefit. Potential to benefit and potential to contribute were rated on a three item (1-2-3) Likert scale while the remaining four areas were rated on a 0–9 Likert scale. The six areas were operationally defined as follows:

Program: The absolute strength of the student's curriculum, considering the number of AP or IB courses taken, the specific AP or IB courses taken, and the extent to which the student has pursued challenging courses across all five of the core academic disciplines (English, math, lab science, social science, foreign language).

Performance: The student's overall academic performance, considering all grades but focusing on performance in the five core academic disciplines—English, math, lab science, social science, and foreign language—and recognizing either upward or downward trends in grades from freshman to senior year.

Extracurricular Activities: The extent to which the student has persisted and achieved in activities outside the classroom, taking into consideration the number of years a student has been involved in a particular activity, the level of achievement, and whether the student has been entrusted with a leadership role or other significant responsibilities.

Essays: The student's ability to communicate effectively in writing and demonstrated likelihood of contributing substantially to the academic environment of the school.

> "Each application was rated by each reader in six different areas– –academic program, academic performance, extracurricular activities, essays, potential to contribute and potential to benefit."

Potential to Contribute: The extent to which the student's unique personal qualities suggest the potential to contribute to the academic or public-service mission of the university and society at large, taking into consideration qualities not adequately reflected in the other ratings, such as a special talent, unusual perspective, cultural background, or set of experiences.

Potential to Benefit: The student's ability to benefit specifically from an education and experience at this university, taking into consideration whether the student will be able to compete and succeed in this environment, his willingness to take advantage of opportunities offered, and the likelihood of achieving beyond the life of university.

Procedures

Fourteen admission counselors working for the same university to which the students applied served as raters. The level of admission experience of the raters ranged from one to eight years.

All raters participated in a two-day training session prior to reading applications. The definitions of the six rating scales were reviewed, sample applications were evaluated by the group and by each individual reader, and discussions followed to clarify rating discrepancies. During the four months following training, the readers evaluated the 17,000 applications received by the university, with each reader responsible for an approximately equal number of applications. Many applications were reviewed by multiple readers and/or discussed by reading committees.

After all applications had been evaluated and decisions finalized, the fourteen raters independently evaluated the 50 applications randomly selected for the sample. Most raters had evaluated at least one of the sample applicants during the reading season, but prior ratings were not made available to any of the raters during the study. The resulting design was a fully-crossed, one-facet design with three sources of variance: persons (p), students who applied to the university; raters (r), admission counselors reading and evaluating the applications; and the persons x rater interaction (pr). Persons, or students, were considered the object of measurement; person variance reflects true differences between students. Raters and persons x raters were considered sources of error variance. Researchers used analysis of variance (ANOVA) procedures to generate variance estimates for each source. G coefficients, or reliability estimates, were calculated for both relative and absolute decisions.

Results

The G Study

Variance estimates for p, r and pr are presented in Table 1. Most of the variance in scores was attributable to individual differences between students for three of the scales: academic performance, academic program and, to a lesser degree, extracurricular activities. The person x rater interaction was the largest source of variance for the remaining three scales, accounting for as much as 62 percent and 57 percent of the total variation in scores for potential to benefit and potential to contribute, respectively. Variance attributable to raters constituted a larger portion of total variance for extracurricular activities and essays than any of the other scales, representing eight percent and 10 percent respectively.

"Persons, or students, were considered the object of measurement; person variance reflects true differences between students. Raters and persons x raters were considered sources of error variance."

The D Studies

Two decision studies were conducted using the variance estimates resulting from the G study. Because admission ratings can potentially be used in two different ways—to judge the relative standing of an applicant compared to other applicants in the pool, or to make a decision about whether or not to admit an applicant—reliability estimates were calculated for both relative and absolute decisions. The dependability

Table 1. Estimated Variance Components by Admission Ratings Scales

	Sources of Variation With Estimated Variance Components and Percent of Total Variance		
Admission Rating Scales	Person (p)	Rater (r)	Person X Rater Interaction (pr)
Academic Program	5.815 (91%)	.008 (0%)	.557 (9%)
Academic Performance	3.901 (89%)	.017 (1%)	.485 (11%)
Extracurricular Activities	1.056 (56%)	.146 (8%)	.670 (36%)
Essays	.6446 (46%)	.134 (10%)	.516 (44%)
Potential to Benefit	1.369 (36%)	.093 (2%)	2.37 (62%)
Potential to Contribute	1.641 (38%)	.219 (5%)	2.502 (57%)

Table 2. Estimated Generalizability Coefficients for Absolute and Relative Decisions

Number of Raters												
	Program		Performance		Activities		Essays		Benefit		Contribute	
	A	R	A	R	A	R	A	R	A	R	A	R
1	.911	.912	.886	.889	.564	.612	.498	.555	.357	.366	.376	.396
2	.954	.954	.940	.941	.721	.759	.665	.714	.527	.536	.547	.567
3	.969	.969	.959	.960	.795	.826	.748	.789	.625	.635	.644	.663
6	.984	.984	.979	.980	.885	.904	.856	.882	.770	.776	.783	.797
10	.990	.991	.987	.988	.928	.940	.908	.926	.848	.852	.868	.868
14	.993	.993	.991	.991	.948	.957	.933	.946	.886	.890	.894	.902
20	.995	.995	.994	.994	.963	.969	.952	.962	.918	.920	.923	.929
40	.998	.998	.997	.997	.980	.980	.975	.980	.957	.959	.960	.963

of scores when making absolute decisions is especially important since the decision to offer or decline admission can significantly impact a student's academic future.

Reliability estimates for each of the six scales are reported in Table 2. The type of decision (relative v. absolute) and number of raters impacted the estimates for each of the six scales, although to different and varying degrees. The reliability estimates for Academic Program and Academic Performance were the most stable, with almost equal values for absolute and relative decisions, and with the least amount of fluctuation as the number of raters was increased from three to 10 (.021 increase for both relative and absolute decisions).

The reliability estimates for extracurricular activities and essays were the least stable, with considerable fluctuation between relative and absolute decisions, and substantial fluctuation with changes in the number of raters. The difference in estimates for relative and absolute decisions can be explained by the larger contribution of rater error. Because rater error impacts everyone equally, relative decisions such as rank ordering are not impacted. Absolute decisions, on the other hand, often rely on a specific cut-off value; such decisions could be significantly impacted by a rater's tendency to rate consistently low. As a result, the estimate for extracurricular activities with three raters was .826 for relative decisions, but only .795 for absolute decisions. Similarly, the estimate for essays with three raters was .789 for relative decisions, and only .748 for absolute.

Varying the number of raters also impacted the reliability estimates of the extracurricular activity and essay scales to a greater magnitude than the other scales. Increasing the number of raters from three to 10 increased the reliability estimate for extracurricular activities to .928 for absolute decisions and .940 for relative; the estimate for essays increased to .908 for absolute and .926 for relative decisions.

The reliability estimates for potential to benefit and potential to contribute also fluctuated as a result of changes in the number of raters. Like academic program and academic performance, however, the reliability estimates were similar for relative and absolute decisions, due to the relatively small contribution of rater error to the overall variance (Table 2).

As a ratio of true score variance (attributable to students) to total variance, it is the magnitude of the reliability estimate that is most important in determining whether it is acceptable to use a score in making decisions about individuals. Standard practice, established by Nunnally (1967), suggests that reliability estimates less than .90 are unacceptable. Using this as a criterion, only the reliability coefficients of academic program and academic performance meet the criterion. It would

"The results of this study suggest that admission personnel should be cautious when using the six ratings scales to make important decisions about applicants—whether in relation to the rest of the applicant pool or in terms of a particular student's admission. Only the reliability estimates for academic program and academic performance reached acceptable levels of reliability. Ten or more readers would be needed to produce reliable ratings for extracurricular activities, essays, potential to benefit, and potential to contribute."

require 10 or more raters to establish an acceptable reliability index for extracurricular activities, essays, potential to benefit, and potential to contribute when making relative decisions about students, and at times, 20 or more raters to achieve a similar level when making absolute decisions.

Discussion

The results of this study suggest that admission personnel should be cautious when using the six ratings scales to make important decisions about applicants—whether in relation to the rest of the applicant pool or in terms of a particular student's admission. Only the reliability estimates for academic program and academic performance reached acceptable levels of reliability. Ten or more readers would be needed to produce reliable ratings for extracurricular activities, essays, potential to benefit, and potential to contribute.

In most admission offices, it is unlikely that applications are read by more than two or three readers; multiple readers create logistical, financial and time-management challenges. For the scales that would require 10 or more readers, administrators may decide to eliminate them from the evaluation process altogether, but they might also consider providing more reader training, improving the operational definitions of the rating scales, and limiting (when possible) the number of inexperienced readers introduced to the evaluation process at any one time. The reliability of the rating scales should be reassessed after training, and arguably, on an annual basis as new readers are introduced to the process and old readers retire.

On a final note, it is important to consider that many admission offices use a holistic approach to reading, such that no single factor alone constitutes a decision to offer or decline admission. Information about program, performance and essays, for example, are used in combination with information about test scores, racial and ethnic background, legacy, and teacher recommendations. In a sense, assessing the reliability of each scale in isolation is an academic exercise more than a practical one; because applicants are compared to one another across a variety of dimensions, and because decisions are based on many sources of information, the reliability of any single scale should perhaps also be placed in a larger context. Low reliability on potential to benefit, for example, does not necessarily suggest low reliability on admission decisions.

Regardless of how rating scales are used in particular admission offices—whether simply as a shorthand way for admission personnel to communicate with one another about applicants or more centrally in the decision-making process—they should not be used without some investment of time. The purposes for which scales are developed, their operational definitions, their reliability, and the practices in place to train readers on how to use them are all aspects of reading that should be regularly evaluated. In short, regular evaluation of the processes used to review applications should become as standard a practice in the profession as evaluating the applicants themselves.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20(1), 37-46.

Hux, K., Sanger, D., Reid, R., & Maschka, A. (1997). Discourse analysis procedures: Reliability issues. *Journal of Communication Disorders* 30, 133-150.

Nunnally, J.C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Suen, H. K. (1990). *Principles of Test Theories.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Swartz, C.W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. L., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement* 59(3), 492-506.

Ueberxsax, J. (2003). *Statistical Methods for Rater Agreement.* [Online]. Available: http://ourworld.compuserve.com/homepages/ jsuebersax/agree.htm